Journal of Experimental Psychology: Applied
2016, Vol. 22, No. 4, 393–405

© 2016 American Psychological Association
1076-898X/16/$12.00 http://dx.doi.org/10.1037/xap0000096

Improving Metacomprehension Accuracy in an Undergraduate Course Context

Jennifer Wiley, Thomas D. Griffin, Allison J. Jaeger, Andrew F. Jarosz, and Patrick J. Cushen
University of Illinois at Chicago

Keith W. Thiede
Boise State University

Students tend to have poor metacomprehension when learning from text, meaning they are not able to distinguish between what they have understood well and what they have not. Although there are a good number of studies that have explored comprehension monitoring accuracy in laboratory experiments, fewer studies have explored this in authentic course contexts. This study investigated the effect of an instructional condition that encouraged comprehension-test-expectancy and self-explanation during study on metacomprehension accuracy in the context of an undergraduate course in research methods. Results indicated that when students received this instructional condition, relative metacomprehension accuracy was better than in a comparison condition. In addition, differences were also seen in absolute metacomprehension accuracy measures, strategic study behaviors, and learning outcomes. The results of the current study demonstrate that a condition that has improved relative metacomprehension accuracy in laboratory contexts may have value in real classroom contexts as well.

Keywords: metacomprehension, comprehension monitoring, expository science text, self-regulated learning

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Success in college depends to a considerable degree on students' ability to engage in effective comprehension of informational text. Even with advances in technology, text still remains a primary source of information transmission in many courses, as students are assigned readings to study on their own and are expected to have the skills necessary to engage in effective self-regulated learning. However, a great deal of research has shown that readers lack an ability to accurately track their comprehension. Metacomprehension accuracy is defined as the ability to predict how well one will do on a test of comprehension after reading. Although this is a critical skill for the regulation of many reading and studying behaviors, empirical studies have demonstrated that most college students are quite poor at gauging how well they have understood what they have just read (Baker, 1989; Pressley, 2000, 2002; for reviews, see Dunlosky & Lipko, 2007; Maki, 1998a; Thiede, Griffin, Wiley, & Redford, 2009). Few students come to college equipped with the monitoring skills they need to engage in maximally effective self-regulated study behaviors that will in turn enable them to successfully comprehend and learn from expository texts (Ley & Young, 1998; Otero & Campanario, 1990; Zimmerman, 2002). Furthermore, it has been shown that as a result of poor metacomprehension accuracy, readers fail to make optimal decisions about what to reread (Maki, 1998a; Rawson, O'Neil, & Dunlosky, 2011; Thiede, Anderson, & Therriault, 2003).

Many models of effective self-regulated learning presume that metacognition, particularly the accurate monitoring of ongoing learning, is what allows for online regulation of cognitive processes during study (e.g., Dunlosky & Metcalfe, 2008; Greene & Azevedo, 2007; Griffin, Wiley, & Salas, 2013; Koriat, 1997; Metcalfe, 2009; Nelson & Narens, 1990; Thiede & Dunlosky, 1999; Winne & Hadwin, 1998; Zimmerman, 2002). As such, accurate monitoring is critical for effective study (Maki, 1998a; Winne & Perry, 2000). If a person does not accurately differentiate well-learned material from less-learned material, he or she could waste time studying material that is already mastered, or worse, fail to restudy material that has not been adequately learned. Thus, to assess whether readers have the ability to differentiate between well-learned and less-learned materials, a paradigm was developed in which participants are given a set of texts or passages to read (Glenberg & Epstein, 1985; Maki & Berry, 1984). Participants then predict their comprehension and complete comprehension tests for each individual passage. Monitoring accuracy is operationalized as the intraindividual correlation between a person's predictive ratings and actual test performance among the set of passages; thus, greater comprehension monitoring accuracy is indexed by a more positive intraindividual correlation. A standard

Jennifer Wiley, Thomas D. Griffin, Allison J. Jaeger, Andrew F. Jarosz, and Patrick J. Cushen, Department of Psychology, University of Illinois at Chicago; Keith W. Thiede, College of Education, Boise State University.

Allison J. Jaeger is now at Department of Psychology, Temple University. Andrew F. Jarosz is now at Department of Psychology, Mississippi State University. Patrick J. Cushen is now at Department of Psychology, Murray State University.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B07460 to Thomas D. Griffin, Keith W. Thiede, and Jennifer Wiley. The opinions expressed are those of the authors and do not represent views of the Institute of Education Sciences or the U.S. Department of Education.

Correspondence concerning this article should be addressed to Jennifer Wiley, Department of Psychology, University of Illinois at Chicago, 1007 W. Harrison Street (MC.285), Chicago, IL 60607. E-mail: jwiley@uic.edu

term for this index is *relative metacomprehension accuracy* (Maki, 1998a). This basic paradigm provides a good model for the actual self-regulated learning that students need to engage in on a daily basis, as students routinely have to study multiple topics simultaneously, within a limited timeframe, and need to regulate how much effort to devote to each topic. Consistent with this assumption, relative metacomprehension accuracy has been demonstrated to relate positively to self-regulated learning outcomes (de Bruin, Thiede, Camp, & Redford, 2011; Rawson et al., 2011; Thiede et al., 2003; Thomas & McDaniel, 2007).

An alternative approach has explored comprehension monitoring accuracy using measures of *absolute accuracy* (i.e., the absolute difference between judgment magnitude and performance), or *confidence bias* (i.e., the signed difference between judgment magnitude and performance). Whereas relative accuracy is measured using intraindividual correlations among a set of judgments and tests for each individual, absolute accuracy and bias measures are computed as difference scores. Absolute accuracy and confidence bias of judgments are not reliably correlated with relative accuracy, and they measure distinct aspects of the metacognitive process (Griffin et al., 2013). If a learner assumes that all tests on all topics will be much easier than they actually are, their judgments will greatly exceed their performance (high overconfidence) and the absolute magnitude of the difference will be high (low absolute accuracy). Yet if they make this same assumption about all the tests, it will have little to no impact on how well the variance in their judgments tracks the variance in their performance across tests (relative accuracy). Reasonably high absolute accuracy and low confidence bias can be achieved by merely making reasonable a priori assumptions about oneself or the tests in general, whereas such assumptions will not help in predicting relative differences in one's performance from text to text. The latter requires more sensitivity to text-specific cues, such as those gleaned from the actual experience of trying to read and comprehend each individual text. In addition to sensitivity to a learner's general a priori assumptions, computations of absolute accuracy and confidence bias are directly dependent on overall test performance itself. This dependence in absolute accuracy and confidence bias measures has been the primary argument used in favor of relative accuracy, as it represents a measure of metacognitive monitoring that is not confounded with test performance (Nelson, 1984).

Most studies that have been done in a classroom context have explored the accuracy of students' monitoring in terms of absolute accuracy, and in general results suggest that students tend to be inaccurate, and often overconfident, in estimating their level of overall mastery of any given topic (Hacker, Bol, & Bahbahani, 2008; Hacker, Bol, & Keener, 2008; Huff & Nietfeld, 2009; Lin & Zabrucky, 1998: Nietfeld, Cao, & Obsorne, 2006; Schraw, 2009). Absolute accuracy and confidence bias are likely to impact study behaviors differently than relative accuracy. As they are heavily determined by the magnitude of a single or aggregated judgment, absolute accuracy and confidence bias would be relevant for decisions to terminate study (Dunlosky & Rawson, 2012). Learners may decide not to study a text any longer when they judge the text to be comprehended, but this decision is not made relative to other texts. So an overestimation of comprehension may lead to an inappropriate early termination of study. On the other hand, relative accuracy is most relevant when learners have a restricted amount of time available to engage in restudy, which forces them to allocate time between texts. That is, relative accuracy should be important for allocating study time and focusing resources where they are most needed when time and resources are limited. Realistically, learners will rarely spend unlimited time to master all to-be-learned materials, so relative accuracy and study-time allocation will be relevant any time more than a single simple concept is being learned. Most prior work exploring metacognitive judgment accuracy in authentic classroom contexts has used measures of absolute accuracy or confidence bias; therefore, less is known about the factors that may affect relative metacomprehension accuracy.

Although this study reports both absolute and relative metacomprehension measures, the main emphasis is in understanding the conditions that might promote better relative accuracy in the context of learning from text for a research methods course. In particular, this study tested whether a condition that included both example comprehension test items on different topics and self-explanation instructions could be shown to improve relative metacomprehension accuracy. Specifically, one half of a research methods class was given a reading instruction that prompted students to attempt to try to explain texts to themselves as they studied, and example comprehension test items on different topics that gave them an expectation that upcoming test items would require them to generate inferences about the readings. Both experimental and control groups of students had the opportunity to read the same set of course-related texts twice, under the same time limitations, that forced them to make study choices. The experimental condition was developed based on research done using a situation-model approach to improving comprehension monitoring, where providing example test items on different topics and explanation prompts have been shown to improve relative monitoring accuracy in laboratory studies, as reviewed below.

The Situation Model Approach to Accurate Metacomprehension

Much of the existing work on *relative* accuracy and regulation of study has explored students' estimates of learning in the context of paired-associate learning; for example, learning foreign language vocabulary, or memory for definitions or facts (Bjork, Dunlosky, & Kornell, 2013; Dunlosky & Rawson, 2012; Koriat, 1997; Metcalfe, 2002; Nelson & Narens, 1990; Son & Metcalfe, 2000; Thiede & Dunlosky, 1999, Vesonder & Voss, 1985). As paired-associate learning involves memorizing items, this work is exploring the construct of meta*memory, or the ability to predict one's own performance on tests of memory.* In this literature, it has been demonstrated that *relative metamemory* accuracy as assessed with intraindividual correlations between predictions of whether or not each item will be remembered, and whether or not each item is remembered successfully, can reach near perfect levels (i.e., intraindividual correlations above .90; Nelson & Dunlosky, 1991; Rhodes & Tauber, 2011). There exists a parallel body of work that has explored students' estimates of learning from text, where students judge the extent to which they have comprehended what they have read. This work explores the construct of meta*comprehension*. In contrast to typical *relative metamemory* accuracy findings, typical levels of *relative metacomprehension* accuracy are quite low. Maki (1998a) reported that the mean intraindividual

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

correlation between comprehension ratings and test performance across 25 studies from her lab was only around .27. Meta-analyses by Dunlosky and Lipko (2007); Lin and Zabrucky (1998); Thiede et al., (2009); and Weaver, Bryant, and Burns (1995) reached similar conclusions. Furthermore, recent work has shown that readers can be easily misled in their monitoring accuracy by font (Miele & Molden, 2010; Sanchez & Jaeger, 2015), by reading from a computer screen instead of paper (Ackerman & Goldsmith, 2011; Lauterman & Ackerman, 2014), by the presence of colorful illustrations or interesting details (Ackerman & Leiser, 2014; Jaeger & Wiley, 2014; Serra & Dunlosky, 2010), by more realistic illustrations (Hegarty, Smallman, & Stull, 2012), and even by analogies that are embedded in science texts with a goal of improving comprehension (Jaeger & Wiley, 2015).

Why are readers so poor at monitoring their own understanding? A prevailing explanation for inaccurate monitoring is that students typically use suboptimal cues to predict their comprehension. Koriat's (1997) cue-utilization framework asserts that learners infer their level of learning and gauge likely future performance on the basis of two general types of cues: those tied to subjective *experiences during learning;* and those tied to *a priori assumptions* about the task, the materials, the topic, the presumed general effectiveness of the study strategies one has employed, or one's abilities. Griffin, Jee, and Wiley (2009) mapped these cue types onto Flavell's (1979) distinction between metaknowledge and metaexperiences, noting that a priori assumptions or metaknowledge provides *heuristic cues* for deriving judgments of understanding, whereas experienced-based cues and metaexperiences represent cues that must be monitored during a learning episode. Only *experience-based cues* directly reflect the level of learning that has actually occurred and the quality of the mental representation one has constructed from reading. This means that experience-based cues will be more *valid* than heuristic cues for predicting the quality of comprehension for particular texts.

However, not all experience-based cues are equally valid indicators of text comprehension. Some cues reflect only surface memory, whereas other cues better reflect understanding of the concepts, their relation to each other, and their relation to prior topic knowledge (Rawson, Dunlosky, & Thiede, 2000; Wiley, Griffin, & Thiede, 2005). Experience-based cues can be mapped onto the different levels of text representation specified in the construction-integration model of text comprehension (Kintsch, 1998). In this model, text is concurrently represented at multiple levels including a lexical or surface level, a textbase level, and a situation model level. The lexical level, containing the surface features of the text, is constructed as the words and phrases appearing in the text are encoded. The textbase level is constructed as segments of the surface text are parsed into propositions, and as links between text propositions are formed based on argument overlap. The construction of the situation model also involves linking propositions. However, the integration of propositions here involves connecting text information with the reader's prior knowledge (McNamara, Kintsch, Songer, & Kintsch, 1996) and making causal connections or inferences (Trabasso & van den Broek, 1985; Wiley & Myers, 2003). It is a person's situation model that largely determines his or her performance on tests of comprehension (Kintsch, 1994, 1998; McNamara et al., 1996; Wiley et al., 2005). Thus, experience-based cues related to the quality of one's situation-model provide the most valid basis for making judgments of text comprehension. The situation model approach to accurate metacomprehension posits that readers need to both *generate* and *select* cues related to their situation models when making comprehension judgments (Griffin et al., 2013). When readers are able to access and use cues related to their situation models, the judgments that readers make about their own understanding will be more accurate.

Manipulations that affect valid cue access. A few manipulations have been shown to make the cues related to the quality of their situation model more *accessible* to college students when making their judgments: delayed generation tasks (keywords, summaries, diagram completion), and self-explanation during rereading. Having students engage in delayed generation of keywords, summaries, or diagrams prior to making judgments has produced levels of relative metacomprehension accuracy that exceed the typical benchmark of .27 in numerous studies (Anderson & Thiede, 2008; de Bruin et al., 2011; Lauterman & Ackerman, 2014; Shiu & Chen, 2013; Thiede & Anderson, 2003; Thiede et al., 2003; Thiede, Dunlosky, Griffin, & Wiley, 2005; Thiede, Griffin, Wiley, & Anderson, 2010; Thiede, Redford, Wiley, & Griffin, 2012; van Loon, de Bruin, van Gog, van Merrienboer, & Dunlosky, 2014). Generating information following reading can yield subjective experience-based cues for the reader about how successfully he or she is able to retrieve information (cf. the modified feedback hypothesis described by Glenberg, Sanocki, Epstein, & Morris, 1987). The timing of some generation tasks is critical, because surface memory for text decays over time, whereas the situation model is robust to such decay (Kintsch, Welsch, Schmalhofer, & Zimny, 1990). Thus, when writing a summary or doing another generative task after a delay, a person will more likely have to rely on their relatively greater access to the situation model of a text than on their quickly decaying surface memory. The delayed generation task will therefore produce cues that can better predict situation-model-level comprehension. Thiede et al. (2010) examined this assumption more directly by having readers report what cues they used when making their judgments following an immediate or delayed keyword generation task. Most readers in both conditions reported relying on how much they could remember. But using memory as a cue was far more predictive of situation-model level comprehension when the keyword generation was at a delay.

More recent work has shown that self-explanation during reading can also allow for access to situation model cues during monitoring. Self-explanation is a question-asking activity that can help learners develop a deeper understanding of material as they study. The goal of self-explanation is for learners to generate statements during study that help to explain and make sense of what they are learning (Chi, 2000). Much work has demonstrated that asking students to engage in constructing explanations or arguments while reading, asking students to answer how-and-why questions, instructing readers to make connections across sentences, or prompting them to consider the meaning and relevance of each sentence to the overall point of the text has been shown to improve inference generation and text comprehension processes (Chi, 2000; King, 1994; McNamara, 2004; Wiley & Voss, 1999). However, aside from improving comprehension itself, the importance of engaging in explanation activities in the present context is that it can help to focus learners on the quality of their situation or causal mental models.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

On the basis of this previous work, Griffin, Wiley, and Thiede (2008) tested whether prompting undergraduate students to explain answers to how-and-why questions might improve relative accuracy. The logic was that attempting to explain a text might generate metaexperiences such as a subjective sense of how hard it is to generate an explanation or how coherent an explanation seems. Participants were randomly assigned to conditions. One group read and self-explained connections between parts of the text as they read a second time. Another group read twice, and a third group read only once. The group prompted to self-explain during rereading demonstrated significantly better relative metacomprehension accuracy. Similar benefits were seen by Jaeger and Wiley (2014) when readers were instructed to self-explain from illustrated texts, and by Redford, Thiede, Wiley, and Griffin (2012), who had middle school students construct concept maps during reading. All these tasks were intended to produce greater access to relevant judgment cues, required generative activity, and required use of one's situation-model text representation. As self-explanation and concept mapping inherently entail use of a reader's situation-model representation, the benefits of these activities on relative metacomprehension accuracy were seen even without a delay between reading and the generative activity.

Manipulations that affect valid cue selection. Another set of findings relates to manipulations designed to help readers to *select* valid cues for comprehension monitoring from among those that are available, for example by giving students the expectancy that their comprehension will be assessed with inference tests rather than memory tests. Students need to understand what it means to "comprehend" an expository text in order to be able to monitor their own comprehension (Wiley, Griffin, & Thiede, 2008). Without specific instructions about what comprehension entails, what their goals for reading should be, and what comprehension tests will be like, students may make monitoring judgments based on memory-based cues, or heuristic cues such as interest, instead of comprehension-based cues. They may read an expository text passively, or with the goal of trying to remember it, rather than with a goal of trying to understand what it is saying. In order to engage in monitoring of the correct behaviors, readers need to appreciate that their goal for reading is to try to understand how or why a phenomenon or process occurs, and that the questions they will be asked will depend on making connections and causal inferences across sentences. Said another way, students need to adopt appropriate *norms* for making study-related decisions (Lipko et al., 2009).

Recent studies have attempted to instill comprehension as a reading goal by using a test-expectancy paradigm. For example, Thiede, Wiley, and Griffin (2011) instructed readers to expect either a memory test or a comprehension test, and then gave them example test items for example passages consistent with these goals. This manipulation resulted in significantly higher relative metacomprehension accuracy for the comprehension expectancy group than the memory expectancy group. Wiley et al. (2008) used a similar manipulation plus a no-expectancy group. Again, the comprehension expectancy condition showed the highest relative metacomprehension accuracy. When comprehension test expectancy was combined with a self-explanation instruction, the benefits to relative monitoring accuracy were found to be additive (Wiley et al., 2008), which supports the theoretical distinction between access and selection components of cue utilization, but also demonstrates that they may be particularly effective when combined. The test-expectancy manipulations just described all used example tests on completely different topics, thus manipulation of comprehension for the target information was not confounded with expectancy, and the studies used a measure of relative accuracy that is not inherently confounded with comprehension effects.

There are other studies that have made reference to the idea of inducing test expectancies and have informed readers about the upcoming tests, but these studies have confounded any expectancy information with direct manipulations of how texts are processed during reading. Thomas and McDaniel (2007) told readers what test to expect, but the test-expectancy was linked to a particular processing manipulation (either a missing letter insertion vs. sentence sorting task) and not manipulated independent from the text-processing manipulation. Another study gave some learners practice and experience with generating and answering their own comprehension questions while reading (Bugg & McDaniel, 2012). However, learners generated comprehension questions while reading the target texts on which metacognitive judgments were later collected, which has been shown to impact comprehension test performance (Davey & McBride, 1986). Because their measure of absolute accuracy is by definition highly confounded with test performance itself, the results are fully explicable by the impact of question generation on comprehension processes directly rather than by effects on monitoring processes. Similarly, Lauterman and Ackerman (2014) gave readers practice with reading and answering test questions on the to-be-learned material and found that this practice had an impact on test performance but not judgment magnitudes. Because they used confidence bias as their measure of monitoring, which is also by definition confounded with test performance, their result is also best accounted for as a direct effect on comprehension rather than monitoring processes. Thus, none of these studies provide evidence for an effect of test expectancies on judgment magnitudes, nor a test of how they may affect relative monitoring accuracy, which is a measure that is orthogonal to overall test performance. Such results are critical to inferring that metacognitive processes have been altered by the manipulation.

From Laboratory to Classroom

The main goal of the present research was to test an instructional condition based in the situation model approach in an authentic learning context, to move our knowledge of its effectiveness beyond the laboratory findings cited above. Few studies have explored relative metacomprehension accuracy in authentic course contexts. Part of the difficulty with doing so is that when readings are too similar in topic, or dependent on each other, it may be impossible to get the independent predictions and test performances that are necessary for computing the relative accuracy measure. Previous work using sections of a single passage or textbook chapter has often resulted in very poor relative accuracy scores (Maki, 1998a; Maki, Foley, Kajer, Thompson, & Willert, 1990; Maki, Jonas, & Kallod, 1994; Ozuru, Kurby, & McNamara, 2012), but it is unclear whether this should be interpreted as poor metacomprehension accuracy on the part of the readers, or as a measurement issue (i.e., a consequence of the necessary dependence between different sections of a single text). Perhaps because

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

of this complication, most work done on metacomprehension accuracy in authentic course contexts has used absolute measures of accuracy or confidence bias that can be computed without apportioning the materials, using only one overall prediction and one test score (Hacker, Bol, & Bahbahani, 2008; Hacker, Bol, & Keener, 2008; Huff & Nietfeld, 2009; Lin & Zabrucky, 1998: Nietfeld et al., 2006; Schraw, 2009).

The present study explored relative metacomprehension accuracy using a set of passages that were on distinct topics taken from the content of a research methods course. All students in a research methods course were asked to study the same set of passages, but only half the students were assigned to an experimental condition that provided them with a self-explanation instruction as well as example tests to encourage comprehension-test expectancy. As a first step in extending effective laboratory manipulations into authentic learning contexts, a main question was whether this condition would result in improved relative metacomprehension accuracy in the context of a research methods course. A second main question was whether any effects would be seen in learning outcomes (quiz scores) or restudy behaviors (selection of which texts to restudy). To date, only a few studies have explored when accurate comprehension monitoring actually translates to better regulation and thus more successful learning and comprehension from text-based materials (de Bruin et al., 2011; Rawson et al., 2011; Thiede et al., 2003; Thiede et al., 2012; Thomas & McDaniel, 2007). To explore this issue, before taking quizzes, all students were given a second opportunity to restudy the passages for a limited time. It was expected that students in the instructional condition might engage in more strategic or more effective study, which should result in more successful learning. Although relative judgment accuracy is not computationally dependent on overall test performance, if readers are making use of their greater accuracy to engage in more optimal restudy, one would expect the two should be correlated and that accuracy should mediate observed performance benefits.

Method

Participants

The sample for this study consisted of students enrolled in a 200-level research methods course (generally taken by second year students, average age 19–20) at a large public university in the United States. Complete data were obtained from 93 participants (43 experimental and 50 control). The sample was 69% female, with no differences between conditions in gender, $X^2 < 1$, and no differences in ACT scores (American College Testing scores, a standardized test of college readiness in the U.S., $M = 24.91$, $SD = 4.04$) between the two conditions, $t < 1$.

Design

The design of the study was between-subjects with two conditions using existing recitation sections (25 students each) led by 3 teaching assistants (TAs), who met 1 day a week to support understanding of concepts presented in the textbook and in the lecture. At the beginning of the semester, each TA was assigned two sections by the instructor, and the TA assignments remained constant throughout the semester. Each TA taught two sections, and one of these sections for each TA was randomly assigned to receive the experimental manipulation while the other section served as a business-as-usual comparison condition.

Materials

Wiley et al. (2005) pointed out that the design of expository texts and comprehension tests are critical for testing metacomprehension accuracy. Models of text processing suggest that comprehension is best represented by a person's situation model, mental model, or causal model of a text (Kintsch, 1994, 1998; Trabasso & van den Broek, 1985; Wiley & Myers, 2003). Only texts that have clearly distinguishable surface and situation-model representations, and only test questions that can be answered using just one or the other representation, will lead to interpretable results for metacomprehension accuracy. In order to create distinct content, target passages were excerpted from different chapters of textbooks, or distant sections that appeared within the same chapter (Gravetter & Forzano, 2008; Stanovich, 2007). The topics for the target texts were Operational Definitions, The Barnum Effect, The Third Variable Problem, The Placebo Effect, and Sampling Bias. The texts were edited so that connections among ideas important for forming a situation model were not fully explicit in the surface form of the text. The texts were between 600 and 750 words in length and were written at a Grade 11–12 level according to Flesch Kincaid. A sample text is included in the Appendix. In addition, there were two practice passages that were included at the start of the first session to familiarize students with making predictive judgments.

Similar considerations are critical for the construction of the comprehension tests. They need to contain more than one or two items (cf. Weaver, 1990) and provide a valid measure of comprehension (i.e., assess the situation model of the text and not just surface memory). This study used multiple-choice tests with four alternatives that asked students to think about possible connections, relations, predictions, or conclusions that were implied by the texts but did not explicitly appear (derived from the meaning-based verification tests used by Royer, Carlo, Dufresne, & Mestre, 1996, and Wiley & Voss, 1999; and similar to tests used by Harp & Mayer, 1998). Five questions were developed for each text. Examples of comprehension test items are included with a sample text in the Appendix. Previous work has shown that performance on these types of questions reliably correlates with other comprehension assessments, including performance on how-and-why essay questions (Hinze, Wiley, & Pellegrino, 2013; Sanchez & Wiley, 2006; Wiley et al., 2009), as well as with performance on ACT tests and the Nelson Denny (Griffin et al., 2008). Test performance in this study correlated with ACT Composite score at $r = .32$, $p < .001$.

Procedure

Students participated in this activity during two class periods at the beginning of the semester as part of their recitation sections. All activities took place through an Internet browser.

For both conditions, the activity in the first session was introduced by this text:

> In this activity, you will practice the study skills that you will need for this course. You will be reading a series of short texts on research

> methods and estimating how many questions you could get correct on a five item multiple-choice comprehension test. You will read and make predictions for each text.

In the experimental condition, the students also received this instruction:

> The literature suggests that people study differently depending on the kind of test they expect. You will be taking tests that assess your ability to make connections between the different parts of a text (i.e., link the parts of the text).
>
> We will start with two texts that will give you practice with these kinds of tests. For those texts, you will get example tests right after you predict your performance.

In the first session, all students read the same set of 2 practice and 5 target passages in the same order. After reading the initial instructions, all participants read and completed predictive judgments for the 2 practice passages. After reading each practice text, participants were asked to predict how many items out of 5 they would be likely to get correct on a quiz. Only the students in the instructional condition then took quizzes on the practice passages as part of the experimental manipulation that set up a comprehension-test expectancy. The students were not given any feedback on their performance on the practice quizzes.

Following the practice phase, participants in the experimental condition received this additional explanation instruction based on Griffin et al. (2008):

> You will now read a second set of texts. As you read each text, you should try to explain to yourself the meaning and relevance of each sentence or paragraph to the overall purpose of the text. Ask yourself questions like:
> - What new information does this paragraph add?
> - How does it relate to previous paragraphs?
> - Does it provide important insights into the major theme of the text?
> - Does the paragraph raise new questions in your mind?
>
> For example, take this paragraph about hail and sleet. Some possible comments you could ask yourself are in quotes:
>
> Sleet are raindrops that freeze on their way down
>
> Hailstones freeze in the cloud then start to fall.
>
> "I wonder what difference that could make?"
>
> Because ice balls are lighter than raindrops, the wind can blow hailstones backup into the clouds.
>
> "What happens when hail goes back into the clouds?"
>
> Water freezes around hailstones again and again in the clouds, until they are heavy enough to reach the ground.
>
> "So that would mean hailstones are usually larger than sleet."
>
> If you look at sleet and hail, hail has many more layers of ice.
>
> "That makes sense if they freeze more than once."
>
> Try your best to think about these issues and ask yourself these kinds of questions about each text as you read. As you finish each paragraph, before you move on to the next paragraph, explain to yourself what that paragraph meant.

The control condition did not receive this instruction. All students in both conditions then had a chance to study the 5 target texts. After reading each text, they predicted their quiz performance for that topic. Study time was not limited, but all students finished reading and making judgments within the 50 min period. Although exact timing measures are not available because of server lags during data collection, the elapsed time from the beginning to the end of the combined study-and-judgment period on the target texts was similar for students in both the control (M = 19.62 min) and experimental (M = 20.24 min) conditions ($t < 1$). Although this is not the most sensitive measure of study time, the lack of differences in elapsed time derived from timestamps suggest that large differences in time on task are not likely to be driving any effects of the manipulation.

During the second session a week later, all students had an opportunity to restudy the 5 target texts for a limited time (5 min). During this study period, students selected texts to restudy from a list of links on a web page with the goal of maximizing their overall quiz scores. They were completely free to read texts in any order they wished, skip texts, return to texts, jump between texts, and so forth. The order that the texts were selected during this restudy period was recorded for later analysis. Immediately after the restudy period, all students took the same quizzes to test their understanding.

Measures

Relative meta comprehension accuracy was computed as an intraindividual Pearson correlation between each participant's predictive judgments and his or her actual test performance (Griffin et al., 2008; Nelson, 1984). Pearson scores were highly correlated with relative accuracy computed with Gamma, $r = .95$, $p < .001$. As per Griffin et al. (2008), Pearson correlations were used in the analyses. As Gamma is only sensitive to differences in direction but not magnitude of variations in judgments versus performance, it results in distributions of coefficients that are ordinal rather than continuous, and non-normally distributed with more extreme maximum values. This makes distributions of Gamma less appropriate when used as a criterion in GLM analyses. Better metacomprehension accuracy is indexed by a stronger intraindividual correlation between test-prediction judgments and inference test performance across the set of texts. In addition, absolute accuracy was also computed in two ways (Maki, 1998b; Schraw, 2009), both as confidence bias (signed difference between predicted performance and actual performance) and absolute error (absolute difference between predicted performance and actual performance). Restudy choices (the order that texts were selected during restudy) were coded as strategic (1) if readers selected specific texts to restudy in any order other than the original order of presentation as listed on the screen. Restudy choices were coded as nonstrategic (0) if students selected to restudy the texts in the original order listed on the screen, or if they failed to select any text for restudy at all.

Results

As shown in Table 1, students in the experimental condition did significantly better on the quizzes, $t(91) = 2.08$, $p < .04$, $d = .43$, 95% CI [.02, .84]. Students in the experimental condition also tended to have lower average judgments of comprehension, but

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 1
Effects of Condition on Judgments, Quiz Performance, and Monitoring Accuracy Measures

	Control	Experimental
Component Measures		
Predictive Judgments (% Correct)	.73 (.15)	.68 (.12)
Quiz Performance (% Correct)	.53 (.13)	.59 (.13)
Monitoring Accuracy Measures		
Relative Accuracy (Pearson *r*)	.08 (.50)	.32 (.44)
Absolute Error (difference score)	1.45 (.53)	1.07 (.50)
Bias (signed difference score)	.98 (.86)	.49 (.78)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

this difference was not significant, $t(91) = 1.40$, $p = .17$. Most students were overconfident, as evidenced by 74.2% of the confidence bias scores being > 0. Therefore, because the manipulation increased average test performance without increasing average judgments, it follows that the manipulation also led to better absolute accuracy in terms of less absolute error, $t(91) = 3.57$, $p < .001$, $d = .74$, 95% CI [.32, 1.16] and lower confidence bias scores, $t(91) = 2.87$, $p < .01$, $d = .60$, 95% CI [.18, 1.01] (lower absolute error scores indicate greater accuracy, whereas lower confidence bias scores indicate less overconfidence).

However, relative accuracy is not a direct byproduct of average judgments and performance. Thus, it provides an independent index of metacognitive effects. Students in the experimental condition were better at monitoring their own comprehension of the readings, as demonstrated by their significantly greater relative accuracy, $t(91) = 2.44$, $p < .02$, $d = .51$, 95% CI [.09, .92]. This suggests they were better able to differentiate the topics that they had understood well from those that they understood less well. Despite their lack of computational dependence, relative accuracy and test performance were expected to correlate, because the restudy sessions gave students the opportunity to use their greater relative accuracy to engage in more efficient regulation to improve their overall learning. In addition to the shared effects of condition on both relative accuracy and test performance, these outcomes were correlated with each other, $r = .29$, $p < .01$, 95% CI [.09, .47].

Also, as shown in Figure 1, a mediation analysis supported relative accuracy as the mediating factor between the manipulation and the test performance effects. The significance of this indirect effect was tested using bootstrapping procedures and the PROCESS macro for SPSS (Hayes, 2013). Unstandardized direct and indirect effects were computed for each of 10,000 bootstrapped samples, and the 95% confidence interval was computed by determining the indirect effects at the 2.5th and 97.5th percentiles. The bootstrapped unstandardized indirect effect via relative accuracy was .41, and the 95% confidence interval excluded zero and ranged from 0.08 to 1.05. In contrast, the direct effect of condition on test performance was .99 with a 95% confidence interval that included zero, −0.35 to 2.33. Thus, the effect of condition on test performance was only significant via the indirect path mediated by relative accuracy. As absolute accuracy and confidence bias are statistical byproducts of average test performance, it would not be meaningful to conduct a parallel mediation analysis with those judgment accuracy measures.

The analysis of restudy patterns revealed that a higher proportion of students in the experimental condition engaged in strategic restudy (58% selected to restudy the texts out of order vs. 42% who selected to restudy texts in order) than in the control condition (34% selected texts out of order vs. 66% who selected to restudy texts in order), X^2 (1, $N = 93$) = 5.44, $p < .02$, $\varphi = .24$, 95% CI [.04, .43]. Furthermore, for those who reread strategically, there were benefits on quiz performance due to the experimental condition (average quiz scores: control = .53 vs. experimental = .61, $t(40) = 2.25$, $p = .03$, $d = .71$, 95% CI [.07, 1.34]), whereas no differences due to condition were seen among students who restudied the texts in order (or not at all; average quiz scores for control = .53 vs. experimental = .56, $t < 1$).

Discussion

Previous laboratory studies have shown that prompting students to engage in self-explanation during study and instilling comprehension-test-expectancy can help to improve relative metacomprehension accuracy. The present results provide evidence that an experimental condition including both of these features can improve students' monitoring of understanding in the context of readings for an undergraduate course in research methods. Consistent with previous research, these results provide additional support for the theoretical and practical value of the situation-model approach to improving metacomprehension accuracy. At the same time, the results are also consistent with previous studies finding lower levels of relative accuracy when text materials are on similar topics, such as portions of a long article or textbook chapter (Maki, 1998a). The overall level of relative accuracy was quite low in this study compared to the levels that have been seen in other studies using a more diverse set of topics. However, despite the possible similarity of the topics in the methods text set and the modest improvements that were seen in relative monitoring accuracy, significant effects were still seen in both study behaviors and in performance on learning measures.

Recently, there has been a great deal of research exploring testing effects and how giving students practice tests on target information can lead to better long-term learning outcomes for that information (Hinze et al., 2013; Jensen, McDaniel, Woodard, & Kummer, 2014; Roediger & Karpicke, 2006). In this literature, researchers have demonstrated that taking practice tests can promote better memory for tested material on a subsequent test of the same material, even when compared to students who spent additional time studying. An important distinction between prior work studying practice tests or repeated testing effects, and the present work on test *expectancy* effects, is that students in the present

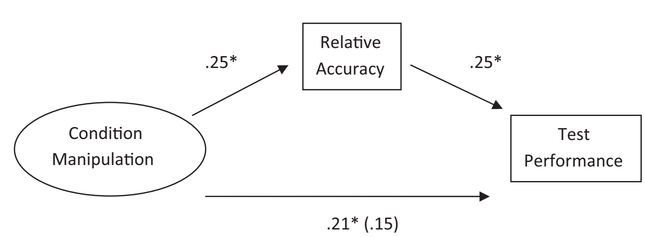

Figure 1. Standardized regression coefficients showing relative accuracy fully mediates the effect of condition on test performance. The direct path controlling for relative accuracy is shown in parentheses. * Confidence interval does not include 0.

study received example test questions on different topics than were used for the target tests, and did not get practice tests for the actual target topics. When studies employ repeated tests on the same materials it is impossible to disentangle whether practice tests may be improving study effectiveness via test expectancy effects, or improving learning via more direct benefits of repeated testing on memory.

Furthermore, when readers are given practice with the exact test items as part of a study session, this fundamentally changes the judgment task from one of *pre*diction to one of *post*diction. Many studies have demonstrated that postdictions can result in more accurate metacomprehension than predictions (Dunlosky, Rawson, & McDonald, 2002; Glenberg & Epstein, 1985; Glenberg et al., 1987; Lin, Moore, & Zabrucky, 2001; Maki et al., 1990; Maki & Serra, 1992a, 1992b; Pierce & Smith, 2001). This has been attributed to the idea that the testing experience offers the reader useful information on their level of understanding even if they are not given explicit feedback on their performance. However, in authentic classroom settings, it is rarely the case that students will be given the same exact test questions in practice tests as in final examinations. Therefore, it remains important to find ways to improve the accuracy of students' *pre*dictive judgments in cases where they do not have access to the exact test questions. The experimental condition used here shows that reader predictions can benefit from instructions that lead them to engage in comprehension processes, to adopt comprehension goals, and to understand what it means to comprehend a text even without prior knowledge of the exact test items. It also means that the current results cannot be explained by "testing effects" because the example tests are on different topics than the target tests.

Few studies have examined both relative and absolute measures of metacomprehension accuracy, and they typically find they are unrelated or affected by different factors (Griffin et al., 2009; Maki, Shields, Wheeler, & Zacchilli, 2005; Schraw, 2009; Thomas & McDaniel, 2007). However, the goal is for students to have both good absolute and relative accuracy. Absolute can determine whether readers persist with studying in general, whereas relative may help them to direct their attention during restudy where it will do the most good. Thus, it is notable that in this study advantages were seen in both measures as a function of the experimental manipulation. Furthermore, it is also notable that the benefits seen in metacomprehension accuracy in this study could be related to differences in study behaviors and in learning outcomes. There are few studies showing that improvements in relative metacomprehension accuracy can relate positively to self-regulated learning outcomes (de Bruin et al., 2011; Rawson et al., 2011; Thiede et al., 2003), and even fewer that measure monitoring judgments after initial study with a restudy opportunity but prior to a final measure of learning. To our knowledge, this study represents the first attempt to show learning improvements due to metacognitive manipulations that are statistically mediated by relative accuracy of monitoring judgments prior to restudy. Perhaps the most important contribution of the present work, however, is showing that instructional manipulations that have improved metacomprehension accuracy in laboratory-based contexts have value in a course-based context as well. At the same time, there are several limitations of the present study that will need to be addressed in the future. For example, there were no measures of prior knowledge for the topics that students were given to study.[1] Furthermore, the instructional manipulation was a compilation of a number of different components including self-explanations, example quiz items, and additional explanations about what to look for while reading. Any of these could have made a difference by themselves. Additional work is still needed to determine which features of this intervention might be most responsible for benefits in comprehension and monitoring accuracy, or if they are working in combination.

[1] Although no measure of prior knowledge for the topics was collected, at the start of the course students were given a short background survey on key statistics concepts (mean, median, mode and variance). The conditions did not differ in their performance on this survey, nor did performance on this survey result in aptitude-by-treatment interactions (ATI). ACT scores were also used to test for ATIs, but none was found for any of the measures.

References

Ackerman, R., & Goldsmith, M. (2011). Metacognitive regulation of text learning: On screen versus on paper. *Journal of Experimental Psychology: Applied, 17,* 18–32. http://dx.doi.org/10.1037/a0022086

Ackerman, R., & Leiser, D. (2014). The effect of concrete supplements on metacognitive regulation during learning and open-book test taking. *British Journal of Educational Psychology, 84,* 329–348. http://dx.doi.org/10.1111/bjep.12021

Anderson, M. C. M., & Thiede, K. W. (2008). Why do delayed summaries improve metacomprehension accuracy? *Acta Psychologica, 128,* 110–118. http://dx.doi.org/10.1016/j.actpsy.2007.10.006

Baker, L. (1989). Metacognition, comprehension monitoring, and the adult reader. *Educational Psychology Review, 1,* 3–38. http://dx.doi.org/10.1007/BF01326548

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology, 64,* 417–444. http://dx.doi.org/10.1146/annurev-psych-113011-143823

Bugg, J. M., & McDaniel, M. A. (2012). Selective benefits of question self-generation and answering for remembering expository text. *Journal of Educational Psychology, 104,* 922–931. http://dx.doi.org/10.1037/a0028661

Chi, M. T. H. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. In R. Glaser (Ed.), *Advances in instructional psychology* (pp. 161–238). Hillsdale, NJ: Erlbaum.

Davey, B., & McBride, S. (1986). Effects of question-generation training on reading comprehension. *Journal of Educational Psychology, 78,* 256–262. http://dx.doi.org/10.1037/0022-0663.78.4.256

de Bruin, A. B., Thiede, K. W., Camp, G., & Redford, J. (2011). Generating keywords improves metacomprehension and self-regulation in elementary and middle school children. *Journal of Experimental Child Psychology, 109,* 294–310. http://dx.doi.org/10.1016/j.jecp.2011.02.005

Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science, 16,* 228–232. http://dx.doi.org/10.1111/j.1467-8721.2007.00509.x

Dunlosky, J., & Metcalfe, J. (2008). *Metacognition*. Thousand Oaks, CA: Sage.

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction, 22,* 271–280. http://dx.doi.org/10.1016/j.learninstruc.2011.08.003

Dunlosky, J., Rawson, K. A., & McDonald, S. L. (2002). Influence of practice tests on the accuracy of predicting memory performance for

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

paired associates, sentences, and text material. In T. Perfect & B. Schwartz (Eds.), *Applied metacognition* (pp. 68–92). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511489976.005

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist, 34,* 906–911. http://dx.doi.org/10.1037/0003-066X.34.10.906

Glenberg, A. M., & Epstein, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 702–718. http://dx.doi.org/10.1037/0278-7393.11.1-4.702

Glenberg, A. M., Sanocki, T., Epstein, W., & Morris, C. (1987). Enhancing calibration of comprehension. *Journal of Experimental Psychology: General, 116,* 119–136. http://dx.doi.org/10.1037/0096-3445.116.2.119

Gravetter, F. J., & Forzano, L. B. (2008). *Research methods for the behavioral sciences* (3rd ed.). Belmont, CA: Wadsworth.

Greene, J. A., & Azevedo, R. (2007). A theoretical review of Winne and Hadwin's model of self-regulated learning: New perspectives and directions. *Review of Educational Research, 77,* 334–372. http://dx.doi.org/10.3102/003465430303953

Griffin, T. D., Jee, B. D., & Wiley, J. (2009). The effects of domain knowledge on metacomprehension accuracy. *Memory & Cognition, 37,* 1001–1013. http://dx.doi.org/10.3758/MC.37.7.1001

Griffin, T. D., Wiley, J., & Salas, C. (2013). Supporting effective self-regulated learning: The critical role of monitoring. In R. Azevedo & V. Aleven (Eds.), *International handbook of metacognition and learning technologies* (pp. 19–34). New York, NY: Springer Science. http://dx.doi.org/10.1007/978-1-4419-5546-3_2

Griffin, T. D., Wiley, J., & Thiede, K. W. (2008). Individual differences, rereading, and self-explanation: Concurrent processing and cue validity as constraints on metacomprehension accuracy. *Memory & Cognition, 36,* 93–103. http://dx.doi.org/10.3758/MC.36.1.93

Hacker, D. J., Bol, L., & Bahbahani, K. (2008). Explaining calibration accuracy in classroom contexts: The effects of incentives, reflection, and explanatory style. *Metacognition and Learning, 3,* 101–121. http://dx.doi.org/10.1007/s11409-008-9021-5

Hacker, D. J., Bol, L., & Keener, M. C. (2008). Metacognition in education: A focus on calibration. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of metamemory and memory* (pp. 429–455). New York, NY: Taylor & Francis.

Harp, S. F., & Mayer, R. E. (1998). How seductive details do their damage: A theory of cognitive interest in science learning. *Journal of Educational Psychology, 90,* 414–434. http://dx.doi.org/10.1037/0022-0663.90.3.414

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York, NY: Guilford Press.

Hegarty, M., Smallman, H. S., & Stull, A. T. (2012). Choosing and using geospatial displays: Effects of design on performance and metacognition. *Journal of Experimental Psychology: Applied, 18,* 1–17. http://dx.doi.org/10.1037/a0026625

Hinze, S. R., Wiley, J., & Pellegrino, J. W. (2013). The importance of constructive comprehension processes in learning from tests. *Journal of Memory and Language, 69,* 151–164. http://dx.doi.org/10.1016/j.jml.2013.03.002

Huff, J. D., & Nietfeld, J. L. (2009). Using strategy instruction and confidence judgments to improve metacognitive monitoring. *Metacognition and Learning, 4,* 161–176. http://dx.doi.org/10.1007/s11409-009-9042-8

Jaeger, A. J., & Wiley, J. (2014). Do illustrations help or harm metacomprehension accuracy? *Learning and Instruction, 34,* 58–73. http://dx.doi.org/10.1016/j.learninstruc.2014.08.002

Jaeger, A. J., & Wiley, J. (2015). Reading an analogy can cause the illusion of comprehension. *Discourse Processes, 52,* 376–405. http://dx.doi.org/10.1080/0163853X.2015.1026679

Jensen, J. L., McDaniel, M. A., Woodard, S. M., & Kummer, T. A. (2014). Teaching to the test . . . or testing to teach: Exams requiring higher order thinking skills encourage greater conceptual understanding. *Educational Psychology Review, 26,* 307–329. http://dx.doi.org/10.1007/s10648-013-9248-9

King, A. (1994). Guiding knowledge construction in the classroom: Effects of teaching children how to question and how to explain. *American Educational Research Journal, 31,* 338–368. http://dx.doi.org/10.3102/00028312031002338

Kintsch, W. (1994). Text comprehension, memory, and learning. *American Psychologist, 49,* 294–303. http://dx.doi.org/10.1037/0003-066X.49.4.294

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York, NY: Cambridge University Press.

Kintsch, W., Welsch, D., Schmalhofer, F., & Zimny, S. (1990). Sentence memory: A theoretical analysis. *Journal of Memory and Language, 29,* 133–159. http://dx.doi.org/10.1016/0749-596X(90)90069-C

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126,* 349–370. http://dx.doi.org/10.1037/0096-3445.126.4.349

Lauterman, T., & Ackerman, R. (2014). Overcoming screen inferiority in learning and calibration. *Computers in Human Behavior, 35,* 455–463. http://dx.doi.org/10.1016/j.chb.2014.02.046

Ley, K., & Young, D. B. (1998). Self-regulation behaviors in underprepared (developmental) and regular admission college students. *Contemporary Educational Psychology, 23,* 42–64. http://dx.doi.org/10.1006/ceps.1997.0956

Lin, L., Moore, D., & Zabrucky, K. M. (2001). An assessment of students' calibration of comprehension and calibration of performance using multiple measures. *Reading Psychology, 22,* 111–128. http://dx.doi.org/10.1080/02702710119125

Lin, L. M., & Zabrucky, K. M. (1998). Calibration of comprehension: Research and implications for education and instruction. *Contemporary Educational Psychology, 23,* 345–391. http://dx.doi.org/10.1006/ceps.1998.0972

Lipko, A. R., Dunlosky, J., Hartwig, M. K., Rawson, K. A., Swan, K., & Cook, D. (2009). Using standards to improve middle school students' accuracy at evaluating the quality of their recall. *Journal of Experimental Psychology: Applied, 15,* 307–318. http://dx.doi.org/10.1037/a0017599

Maki, R. H. (1998a). Test predictions over text material. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 117–144). Hillsdale, NJ: LEA.

Maki, R. H. (1998b). Metacomprehension of text: Influence of absolute confidence level on bias and accuracy. In D. L. Medin (Ed.), *The psychology of learning and motivation: Advances in research and theory* (pp. 223–248). San Diego, CA: Academic Press. http://dx.doi.org/10.1016/S0079-7421(08)60188-7

Maki, R. H., & Berry, S. L. (1984). Metacomprehension of text material. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 663–679. http://dx.doi.org/10.1037/0278-7393.10.4.663

Maki, R. H., Foley, J. M., Kajer, W. K., Thompson, R. C., & Willert, M. G. (1990). Increased processing enhances calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 609–616. http://dx.doi.org/10.1037/0278-7393.16.4.609

Maki, R. H., Jonas, D., & Kallod, M. (1994). The relationship between comprehension and metacomprehension ability. *Psychonomic Bulletin & Review, 1,* 126–129. http://dx.doi.org/10.3758/BF03200769

Maki, R. H., & Serra, M. (1992a). The basis of test predictions for text material. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 116–126. http://dx.doi.org/10.1037/0278-7393.18.1.116

Maki, R. H., & Serra, M. (1992b). Role of practice tests in the accuracy of test predictions on text material. *Journal of Educational Psychology, 84,* 200–210. http://dx.doi.org/10.1037/0022-0663.84.2.200

Maki, R. H., Shields, M., Wheeler, A. E., & Zacchilli, T. L. (2005). Individual differences in absolute and relative metacomprehension accuracy. *Journal of Educational Psychology, 97,* 723–731. http://dx.doi.org/10.1037/0022-0663.97.4.723

McNamara, D. S. (2004). SERT: Self-explanation reading training. *Discourse Processes, 38,* 1–30. http://dx.doi.org/10.1207/s15326950dp3801_1

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14,* 1–43. http://dx.doi.org/10.1207/s1532690xci1401_1

Metcalfe, J. (2002). Is study time allocated selectively to a region of proximal learning? *Journal of Experimental Psychology: General, 131,* 349–363. http://dx.doi.org/10.1037/0096-3445.131.3.349

Metcalfe, J. (2009). Metacognitive judgments and control of study. *Current Directions in Psychological Science, 18,* 159–163. http://dx.doi.org/10.1111/j.1467-8721.2009.01628.x

Miele, D. B., & Molden, D. C. (2010). Naive theories of intelligence and the role of processing fluency in perceived comprehension. *Journal of Experimental Psychology: General, 139,* 535–557. http://dx.doi.org/10.1037/a0019745

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95,* 109–133. http://dx.doi.org/10.1037/0033-2909.95.1.109

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed JOL effect." *Psychological Science, 2,* 267–270. http://dx.doi.org/10.1111/j.1467-9280.1991.tb00147.x

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation* (pp. 125–173). New York, NY: Academic Press. http://dx.doi.org/10.1016/S0079-7421(08)60053-5

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2006). The effect of distributed monitoring exercises and feedback on performance, monitoring accuracy, and self-efficacy. *Metacognition and Learning, 1,* 159–179. http://dx.doi.org/10.1007/s10409-006-9595-6

Otero, J. C., & Campanario, J. M. (1990). Comprehension evaluation and regulation in learning from science texts. *Journal of Research in Science Teaching, 27,* 447–460. http://dx.doi.org/10.1002/tea.3660270505

Ozuru, Y., Kurby, C. A., & McNamara, D. S. (2012). The effect of metacomprehension judgment task on comprehension monitoring and metacognitive accuracy. *Metacognition and Learning, 7,* 113–131. http://dx.doi.org/10.1007/s11409-012-9087-y

Pierce, B. H., & Smith, S. M. (2001). The postdiction superiority effect in metacomprehension of text. *Memory & Cognition, 29,* 62–67. http://dx.doi.org/10.3758/BF03195741

Pressley, M. (2000). What should comprehension instruction be the instruction of? In M. Kamil, P. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research* (pp. 545–561). Mahwah, NJ: Erlbaum.

Pressley, M. (2002). Metacognition and self-regulated comprehension. In A. Farstrup & S. J. Samuels (Eds.), *What research has to say about reading instruction* (3rd ed., pp. 184–200). Newark, DE: International Reading Association.

Rawson, K. A., Dunlosky, J., & Thiede, K. W. (2000). The rereading effect: Metacomprehension accuracy improves across reading trials. *Memory & Cognition, 28,* 1004–1010. http://dx.doi.org/10.3758/BF03209348

Rawson, K. A., O'Neil, R., & Dunlosky, J. (2011). Accurate monitoring leads to effective control and greater learning of patient education materials. *Journal of Experimental Psychology: Applied, 17,* 288–302. http://dx.doi.org/10.1037/a0024749

Redford, J. S., Thiede, K. W., Wiley, J., & Griffin, T. D. (2012). Concept mapping improves metacomprehension accuracy among 7th graders. *Learning and Instruction, 22,* 262–270. http://dx.doi.org/10.1016/j.learninstruc.2011.10.007

Rhodes, M. G., & Tauber, S. K. (2011). The influence of delaying judgments of learning on metacognitive accuracy: A meta-analytic review. *Psychological Bulletin, 137,* 131–148. http://dx.doi.org/10.1037/a0021705

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17,* 249–255. http://dx.doi.org/10.1111/j.1467-9280.2006.01693.x

Royer, J. M., Carlo, M. S., Dufresne, R., & Mestre, J. (1996). The assessment of levels of domain expertise while reading. *Cognition and Instruction, 14,* 373–408. http://dx.doi.org/10.1207/s1532690xci1403_4

Sanchez, C. A., & Jaeger, A. J. (2015). If it's hard to read, it changes how long you do it: Reading time as an explanation for perceptual fluency effects on judgment. *Psychonomic Bulletin & Review, 22,* 206–211. http://dx.doi.org/10.3758/s13423-014-0658-6

Sanchez, C. A., & Wiley, J. (2006). An examination of the seductive details effect in terms of working memory capacity. *Memory & Cognition, 34,* 344–355. http://dx.doi.org/10.3758/BF03193412

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning, 4,* 33–45. http://dx.doi.org/10.1007/s11409-008-9031-3

Serra, M. J., & Dunlosky, J. (2010). Metacomprehension judgements reflect the belief that diagrams improve learning from text. *Memory, 18,* 698–711. http://dx.doi.org/10.1080/09658211.2010.506441

Shiu, L. P., & Chen, Q. (2013). Self and external monitoring of reading comprehension. *Journal of Educational Psychology, 105,* 78–88. http://dx.doi.org/10.1037/a0029378

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 204–221. http://dx.doi.org/10.1037/0278-7393.26.1.204

Stanovich, K. (2007). *How to think straight about psychology* (8th ed.). Boston, MA: Pearson.

Thiede, K. W., & Anderson, M. C. M. (2003). Summarizing can improve metacomprehension accuracy. *Contemporary Educational Psychology, 28,* 129–160. http://dx.doi.org/10.1016/S0361-476X(02)00011-5

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95,* 66–73. http://dx.doi.org/10.1037/0022-0663.95.1.66

Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 1024–1037. http://dx.doi.org/10.1037/0278-7393.25.4.1024

Thiede, K. W., Dunlosky, J., Griffin, T. D., & Wiley, J. (2005). Understanding the delayed-keyword effect on metacomprehension accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 1267–1280. http://dx.doi.org/10.1037/0278-7393.31.6.1267

Thiede, K. W., Griffin, T. D., Wiley, J., & Anderson, M. C. M. (2010). Poor metacomprehension accuracy as a result of inappropriate cue use. *Discourse Processes, 47,* 331–362. http://dx.doi.org/10.1080/01638530902959927

Thiede, K. W., Griffin, T. D., Wiley, J., & Redford, J. S. (2009). Metacognitive monitoring during and after reading. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition and self-regulated learning* (pp. 85–106). Mahwah, NJ: Erlbaum.

Thiede, K. W., Redford, J. S., Wiley, J., & Griffin, T. D. (2012). Elementary school experience with comprehension testing may influence meta-

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

comprehension accuracy among seventh and eighth graders. *Journal of Educational Psychology, 104,* 554–564. http://dx.doi.org/10.1037/a0028660

Thiede, K. W., Wiley, J., & Griffin, T. D. (2011). Test expectancy affects metacomprehension accuracy. *British Journal of Educational Psychology, 81,* 264–273. http://dx.doi.org/10.1348/135910710X510494

Thomas, A. K., & McDaniel, M. A. (2007). Metacomprehension for educationally relevant materials: Dramatic effects of encoding-retrieval interactions. *Psychonomic Bulletin & Review, 14,* 212–218. http://dx.doi.org/10.3758/BF03194054

Trabasso, T., & van den Broek, P. (1985). Causal relatedness and importance of story events. *Journal of Memory and Language, 24,* 595–611. http://dx.doi.org/10.1016/0749-596X(85)90048-8

van Loon, M. H., de Bruin, A. B., van Gog, T., van Merriënboer, J. J., & Dunlosky, J. (2014). Can students evaluate their understanding of cause-and-effect relations? The effects of diagram completion on monitoring accuracy. *Acta Psychologica, 151,* 143–154. http://dx.doi.org/10.1016/j.actpsy.2014.06.007

Vesonder, G. T., & Voss, J. F. (1985). On the ability to predict one's own responses while learning. *Journal of Memory and Language, 24,* 363–376. http://dx.doi.org/10.1016/0749-596X(85)90034-8

Weaver, C. A. (1990). Constraining factors in calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 214–222. http://dx.doi.org/10.1037/0278-7393.16.2.214

Weaver, C. A., Bryant, D. S., & Burns, K. D. (1995). Comprehension monitoring: Extensions of the Kintsch and van Dijk model. In C. A. Weaver, S. Mannes, & C. Fletcher (Eds.), *Discourse comprehension: Essays in honour of Walter Kintsch* (pp. 177–193). Hillsdale, NJ: Erlbaum.

Wiley, J., Goldman, S., Graesser, A., Sanchez, C. A., Ash, I. K., & Hemmerich, J. (2009). Source evaluation, comprehension, and learning in internet science inquiry tasks. *American Educational Research Journal, 46,* 1060–1106. http://dx.doi.org/10.3102/0002831209333183

Wiley, J., Griffin, T., & Thiede, K. W. (2005). Putting the comprehension in metacomprehension. *The Journal of General Psychology, 132,* 408–428. http://dx.doi.org/10.3200/GENP.132.4.408-428

Wiley, J., Griffin, T. D., & Thiede, K. W. (2008). To understand your understanding one must understand what understanding means. *Proceedings of the 30th Annual Conference of the Cognitive Science Society*.

Wiley, J., & Myers, J. L. (2003). Availability and accessibility of information and causal inferences from scientific text. *Discourse Processes, 36,* 109–129. http://dx.doi.org/10.1207/S15326950DP3602_2

Wiley, J., & Voss, J. F. (1999). Constructing arguments from multiple sources: Tasks that promote understanding and not just memory for text. *Journal of Educational Psychology, 91,* 301–311. http://dx.doi.org/10.1037/0022-0663.91.2.301

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 277–304). Mahwah, NJ: Erlbaum.

Winne, P. H., & Perry, N. E. (2000). Measuring self-regulated learning. In M. Boekaerts & P. Pintrich (Eds.), *Handbook of self-regulation* (pp. 531–566). New York, NY: Academic Press. http://dx.doi.org/10.1016/B978-012109890-2/50045-7

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice, 41,* 64–70. http://dx.doi.org/10.1207/s15430421tip4102_2

(*Appendix follows*)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Appendix

Sample Research Methods Passage and Comprehension Questions

The Third Variable Problem

Several years ago, a large-scale study of the factors relating to the use of contraceptive devices was conducted in Taiwan. A large research team of social scientists and physicians collected data on a wide range of behavioral and environmental variables. The researchers were interested in seeing what variables best predicted the use of birth control methods. After collecting the data, they found that the one variable most strongly related to contraceptive use was the number of electrical appliances (toasters, fans, etc.) in the home.

This result probably does not tempt you to propose that the teenage pregnancy problem should be dealt with by passing out free toasters in high schools. But why aren't you tempted to think so? The correlation between appliances and contraceptive use was indeed strong, and this variable was the single best predictor among the many variables that were measured. Your inclination may be to think that it is not the strength but the nature of the relationship that is relevant. Starting a free toaster program would imply the belief that possession of toasters *causes* people to use contraceptives. Identifying a correlation between two variables is not enough to demonstrate that one variable causes changes in the other.

One reason why we cannot infer causation from correlation is that correlational research is vulnerable to the *third-variable problem*. Variables do not exist in isolation, but are actually part of a large and tangled network of interrelated variables. Under these circumstances, changes in one variable are typically accompanied by changes in many other related variables. Therefore, it is possible for relationships to exist between two measured variables not because one causes changes in the other, but because both are commonly influenced by an unmeasured third variable. For example, parents' income could influence both the number of costly appliances in the house and the likelihood of using contraception.

Fortunately, there exist complex correlational statistics that are designed to address problems such as this one. These statistics allow the relationship between two variables to be recalculated after the influences of other variables have been removed. However, to do this correctly, the third variable must be assessed at the same time as the two original variables. For each "third" variable that is measured and included in analyses to statistically remove its influence, we increase the plausibility that the correlation reflects a real direct causal effect between the two original variables.

Our ability to recognize the toaster-as-contraception program as an absurd suggestion is because we have existing beliefs and assumptions about contraception and toasters that make their causal connection seem silly. Unfortunately, the limitations of correlational evidence are not always so easy to recognize when we don't happen to already have relevant and accurate knowledge to prevent us from making the leap from correlation to causation.

Consider, for example, the extremely popular hypothesis in education and social services research that school achievement problems, drug abuse, teenage pregnancy, and many other problem behaviors are the result of low self-esteem. Correlational research frequently identifies a relationship between self-esteem and these behavioral variables. Existing theories about the causal importance of self-esteem lead some people to conclude that low self-esteem leads to problem behaviors, and high self-esteem leads to high educational achievement and accomplishments in other domains. This assumption of causal direction provided the motivation for many educational programs emphasizing increasing self-esteem. Although some of this research attempts to statistically reduce the influence of third variables, such as socio-economic status, it almost entirely fails to consider the *directionality problem.*

In reality, the relationship between self-esteem and school achievement could run in the opposite direction. Superior accomplishment in school and other aspects of life probably lead to higher self-esteem, instead of the other way around. The directionality problem highlights the fact that, before jumping to the conclusion that a correlation between variable A and variable B is due to changes in A causing changes in B, we must first recognize that the direction of causation may be the opposite, from B to A.

These ambiguities prevent us from inferring or interpreting a causal relationship from merely having observed that two variables are correlated. In order to strongly demonstrate causality, it is necessary to engage in experimental manipulation. If manipulating variable A consistently results in changes in variable B, then we can be confident that there is a causal relationship between A and B.

1. Which of the following are TRUE regarding our prior assumptions and interpreting correlational data?

   a) We are less likely to assume a causal relationship from a correlation if we have no prior assumptions at all about the topic or the variables.

   b) Our prior assumptions can sometimes bias us towards assuming there is a causal relationship.

   c) Both of the above

   d) None of the above

(*Appendix continues*)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

2. Which one of the following statements about ways to deal with the problems regarding interpretation of correlations is TRUE?

   a) Complex statistics can be used to address both the third variable and directionality problem.

   b) Experimental manipulations solve the directionality problem, but not the third variable problem.

   c) Experimental manipulations solve the third variable problem, but not the directionality problem.

   d) Using statistics to address the third variable problem requires that the researcher to be aware of the possible third variables before conducting the research.

3. Imagine that we observe a correlation between variables A and B. If we could measure and statistically eliminate the influence of all possible other variables that might cause both A and B, then we could . . .

   a) conclude that A causes B.

   b) greatly reduce the third variable problem.

   c) Both of the above

   d) None of the above

4. A researcher finds that income and happiness are correlated and concludes that more income causes more happiness. Which of the following is an example of a third variable explanation for this correlation?

   a) Happiness causes higher incomes because it motivates hard work.

   b) Income causes happiness because it makes one more attractive to the opposite sex.

   c) Greater education makes people happier and also increases their income.

   d) None of the above

5. Why does the third variable problem exist?

   a) because of the directionality problem

   b) because researchers do not carefully record their measurements

   c) because any 2 variables that are correlated probably both cause each other

   d) because most variables cause changes in many others

Received March 7, 2016
Revision received June 27, 2016
Accepted June 29, 2016 ■

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.